修士論文 要旨

部分事後予測分布に基づくモデル検証の漸近的性質

数理情報学専攻 076231　村岡優輔

指導教員　駒木文保 准教授

1　はじめに

モデル検証とは，統計モデルを用いてデータを解析する際，データに統計モデルが適合していないならばそれを検出するための手法のことを指す．モデル検証は解析の初期で，統計モデルを試行錯誤しながら設定する段階で，仮定したモデルの問題点を発見することを目的とする．そのために，問題点に対応する統計量 T を用意し，観測された値が大きければ問題と判断する．

問題と判断する値の基準に p 値を用いることとする．モデルが正しいとしたとき，p 値は $[0,1]$ 上の一様分布に従うことが求められる．単純にパラメータを推定した場合にはこの性質が壊れることが指摘されており，部分事後予測分布を用いることが提案された [1]．一般的な性質も，[2] によって $O_p(1)$ の評価で示されている．しかし，部分事後予測分布を計算する際，どのような事前分布を選ぶべきかは明らかにされていない．

本発表では，部分事後予測分布による p 値を漸近展開することによって選ぶべき事前分布を調べた結果を説明する．

2　モデル検証と部分事後予測分布による p 値

2.1　モデル検証における p 値の問題

モデル検証において p 値を求める際，モデルは確率分布の集合であるため，どのような確率分布を用いるか自明ではない．χ^2 統計量による検定のように，パラメータに依存しない漸近分布を持つ統計量を用いる場合には問題がないが，ここではそうでない統計量に対しても p 値を計算することを目的とする．

2.2　部分事後予測分布による p 値

部分事後予測分布による p 値を定義する．ここから検証したいモデルの確率密度関数を $f(x;u)$(u がモデルのパラメータ)，検定統計量を $T=t(X)$，X,T の観測値を $x_{\mathrm{obs}}, t_{\mathrm{obs}}$ で表すこととする．

事前分布 $\pi(u)$ を定める．このとき部分事後予測分布による p 値 $p_{\mathrm{ppp}}(x_{\mathrm{obs}})$ は以下で表される．

$$p_{\mathrm{ppp}}(x_{\mathrm{obs}}) := \int 1_{\{t(x)\geq t_{\mathrm{obs}}\}}(x) f(x;u)\pi(u|x_{\mathrm{obs}}\setminus t_{\mathrm{obs}})\mathrm{d}u\mathrm{d}x,$$

ただし，$\pi(u|x_{\mathrm{obs}}\setminus t_{\mathrm{obs}})$ は，T に関する条件付き分布の事後分布である以下の分布である．

$$\pi(u|x_{\mathrm{obs}}\setminus t_{\mathrm{obs}}) := \frac{f(x_{\mathrm{obs}}|t_{\mathrm{obs}},u)\pi(u)}{\int f(x_{\mathrm{obs}}|t_{\mathrm{obs}},u')\pi(u')\mathrm{d}u'}$$

条件付き分布を用いることで，T に関する情報を除いて u を推定することを意図している．

3　漸近展開

モデルが $(m+1,m)$ 次元曲指数型分布族であり，検定統計量が指数型分布族の十分統計量の滑らかな関数である場合を調べた．[3] の推定量の漸近展開を参考にしている．発表では，検定統計量の期待値がモデルのもとで 0 である場合について説明する．指数型分布族の確率密度関数を $g(x;u,t)$ で表す．$t=0$ で仮定したモデルを表すこととする．

まず，観測値が固定された元での $p_{\mathrm{ppp}}(x_{\mathrm{obs}})$ を近似した．予測分布の漸近展開については [4] を参考にした．そこで得られた $p_{\mathrm{ppp}}(x_{\mathrm{obs}})$ が，$x_{\mathrm{obs}}\sim f(x;u_0)=g(x;u_0,0)$ のもとでどのような挙動を示すか調べた．

得られた $p_{\mathrm{ppp}}(x_{\mathrm{obs}})$ の展開より，$p_{\mathrm{ppp}}(x_{\mathrm{obs}})=1-\Phi(\tilde{t^*}_{\mathrm{obs}})+o(N^{-1})$ なる $\tilde{T^*}$ を求めた．$\tilde{T^*}$ は事前分布に依存する．$\tilde{T^*}$ が標準正規分布に従うならば，$p_{\mathrm{ppp}}(x_{\mathrm{obs}})$ が一様分布に従う．$\tilde{T^*}$ が標準正規分布に従う条件として，以下を得る．

定理 3.1

π^* を，事前分布を Jeffreys prior で割った量として次で定義する．

$$\pi^*(u)=\frac{\pi(u)}{|g_{ab}(u)|^{\frac{1}{2}}}$$

このとき，$x_{\mathrm{obs}}\sim f(x;u_0)$ のもとで，p 値が一様分布に従う十分条件は，以下で書ける．アインシュタイン規約を用いている．

$$\begin{aligned}\big\{&(\partial_a\partial_b\log(g_{tt}(u_0)))-\overset{0}{\Gamma}{}_{abc}(u_0)g^{cd}(u_0)\partial_d\log(g_{tt}(u_0))\\&+(\partial_a\log(g_{tt}(u_0)))(\partial_b\log(\pi^*(u_0)))\big\}g^{ab}(u_0)=0\end{aligned}\tag{1}$$

ただし，g_{ab}，g_{tt} は Fisher 情報行列であり，$g^{ab}(u)$ は

$g_{ab}(u)$ の逆行列，$\Gamma^0{}_{abc}(u)$ は接続係数である．

$$g_{\alpha\beta}(u) := E[\partial_\alpha \log g(X;u,t)\partial_\beta \log g(X;u,t)]$$
$$\overset{0}{\Gamma}{}_{abc}(u) := E\Big[\Big\{\partial_a\partial_b \log g(X;u,t) + \frac{1}{2}\partial_a \log g(X;u,t)\partial_b \log g(X;u,t)\Big\} \partial_c \log g(X;u,t)\Big]$$

以下のことが分かった．

- $O(N^{-\frac{1}{2}})$ の評価では，事前分布に依存せず一様分布に従っていると言える．
- $O(N^{-1})$ の評価では，式 (1) を満たす π^* により π を定めれば，一様分布に従うと言える．
- 式 (1) の条件は，$g_{tt}(u)$ により書かれている．これは，事前分布を検証統計量の性質に応じて選ばなければならないことを意味する．

4 具体例における p 値の漸近展開と，適切な事前分布

次のような階層モデルを考える．ただし σ_0^2 は既知とする．

$$\begin{aligned} X_{1i} &\overset{\text{i.i.d.}}{\sim} \mathrm{N}(\theta_{1i}, \sigma_0^2) \\ X_{2j} &\overset{\text{i.i.d.}}{\sim} \mathrm{N}(\theta_{2j}, \sigma_0^2) \\ \theta_{1i} &\overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu, \tau^2) \quad i = 1, \ldots, N \\ \theta_{2j} &\overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu, \tau^2) \quad j = 1, \ldots, N \end{aligned}$$

検定統計量を次のように置く．

$$T := \Big|\frac{1}{N}\sum_{i=1}^{N} X_{1i} - \frac{1}{N}\sum_{j=1}^{N} X_{2j}\Big|$$

この検証統計量はモデルのもとで期待値がパラメータに依存する．しかし，$T' = \frac{1}{N}\sum_{i=1}^{N} X_{1i} - \frac{1}{N}\sum_{j=1}^{N} X_{2j}$ を用いた p 値を $p'(x_{\text{obs}})$ とすると，$p(x_{\text{obs}}) = 1 - 2|p'(x_{\text{obs}}) - \frac{1}{2}|$ と書ける．よって $p'(x_{\text{obs}})$ が一様分布に従えば，$p(x_{\text{obs}})$ も一様分布に従う．T' はモデルのもとで期待値が 0 であるので，$p'(x_{\text{obs}})$ が一様分布に従う条件は式 (1) により得られ，それに基づき事前分布を選択することができる．

式 (1) より， 事前分布を

$$\pi(\mu, \tau^2) = (\tau^2 + \sigma_0^2)^{-1}$$

と選ぶと，部分事後予測分布による p 値は $O_p(N^{-1})$ の評価で一様分布に従う．$(\mu_0, \tau_0^2) = (10, 10)$ のもとで発生させたサンプルに対して部分事後予測分布による p 値とその近似を計算し，その経験累積分布関数を示す．図 1 に，事前分布を $\pi_J(\mu, \tau^2) = (\tau^2 + \sigma_0^2)^{-\frac{3}{2}}$(Jeffreys' prior)，$\pi_r(\mu, \tau^2) = (\tau^2 + \sigma_0^2)^{-1}$ としたときの結果でを示す． グループ数が少ないとき，事前分布を π_r として

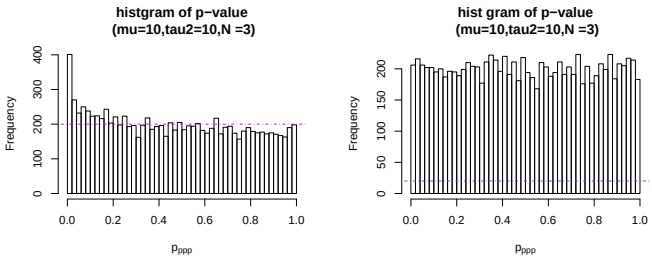

図 1. 部分事後予測分布による p 値のヒストグラム (繰りかえし 10000 回，$N = 3$，左：事前分布 π_J，右：事前分布 π_r)

計算した p 値のほうがより一様分布に近いことが見て取れる．

5 結論と今後の課題

モデルが曲指数型分布族であり，統計量が十分統計量の滑らかな関数である場合について，部分事後予測分布による p 値の漸近展開を導出した．その結果を用いて，事前分布を検定統計量に応じて定めることにより，より一様分布に近い分布に従う p 値を計算することが可能となった．

今後の課題としては，まず，式 (1) を満たす事前分布を求める方法を見つけることが求められる．また，統計量としてサンプルの最大値，最小値を用いる場合がよく研究されているため，その場合にも本研究と対応する結果を導くことが必要と考える．

参考文献

[1] M. J. Bayarri and J. O. Berger. p values for composite null models. *Journal of the American Statistical Association*, 95: 1127–1142, 2000.

[2] J. M. Robins, A. van der Vaart, and V. Ventura. Asymptotic distribution of p values in composite null models. *Journal of the American Statistical Association*, 95: 1143–1156, 2000.

[3] S. Amari. *Differential-Geometrical Methods in Statistics*. Springer, New York, 1985.

[4] F. Komaki. On asymptotic propertiees of predictive distributions. *Biometrika*, 83: 299–313, 1996.